UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Geociências e Ciências Exatas - IGCE
Curso de Bacharelado em Ciências da Computação

MAICON DALL'AGNOL

TRABALHO DE REGRAS DE ASSOCIAÇÃO

Professora: Dra. Adriane Beatriz de Souza Serapião

Rio Claro - SP

2019

1 Introdução

Este trabalho consiste em aplicar o conhecimento de regras de associação adquirido na disciplina Tópicos: Aprendizado de Máquina, tendo assim como objetivo:

- Escolha dois datasets específicos para a tarefa de Regras de Associação.
- Para cada dataset, você irá aplicar os algoritmos APriori e FP-Growth. Seu objetivo é minerar regras que contenham informações relevantes no dataset, seja porque alguma combinação de itens aparece com muita frequência, seja porque alguma combinação de itens não aparece com muita frequência mas está se destacando. Para isso, você deve variar os parâmetros de suporte, confiança e lift.
- Existem duas métricas não estudadas em sala de aula, 'leverage' e 'conviction', que estão disponibilizadas no pacote mlxtend para regras de associação. Estude essas métricas e explique como elas podem contribuir para as análises dos seus datasets.
- Compare os resultados gerados pelos dois algoritmos. Conclua sobre as diferenças encontradas nos resultados de cada dataset na aplicação dos dois algoritmos.

2 Desenvolvimento

Para o desenvolvimento das atividades inicialmente foram escolhidos duas bases bases de dados. A primeira base a ser utilizada corresponde a dados de *reviews* de um E-Commerce de Roupas Femininas contendo informações como idade, avaliação, categoria que o produto pertence, entre outras; A segunda base é composta dados educacionais que é coletado do sistema de gerenciamento de aprendizado, nela há dados como genero, nacionalidade, número de vezes que o aluno levanta a mão, entre outros dados.

2.1 Pré-processamento e Visualização

Ambas bases haviam dados categóricos e numéricos, na primeira base alguns atributos numéricos foram removidos e outros foram discretizados, ainda na primeira base também foram removidas transações que continham item (atributos=valor) faltantes; Para a segunda base visualizou através de um *boxplot* que os dados numéricos estavam variando de 0 a 100 com uma média diferente para cada atributo, desta forma todos os dados numéricos foram discretizados em categorias binárias (abaixo ou acima média do atributo).

Para a aplicação dos algoritmos também transformou-se os dados do formato dataframe para um array da biblioteca Numpy.

2.2 Regras de Associação

Para aplicação do algoritmo Apriori inicialmente transformou-se os dados para um formato transacional utilizando TransactionEncoder, dessa forma utilizou-se o algoritmo apriori implementado na biblioteca mlxtend para extração dos *itemsets* frequentes e $association_rules$ para extração das regras com confiança acima do corte.

Para aplicação Fp-Growth primeiramente implementou-se duas funções cujo propósito era: a. verificação de suporte e b. aplicação, cálculos das medidas. Nesta etapa algumas dificuldades foram encontradas uma vez que a implementação do algoritmo fp_growth baixada pelo indexador de pacotes de python PyPI não retornava todas as medidas, portanta algumas tiveram que ser calculadas o que gerou alguns problemas ao se buscar os suportes do antecedente e consequente.

2.3 Avaliação